ZUSAMMENHÄNGE ZWISCHEN METRISCHEN VARIABLEN III

Martin-Luther-Universität Halle-Wittenberg
Institut für Soziologie
Übung Einführung in die deskriptive Statistik

Agenda

- Wiederholung
- Vertiefung bivariate lineare Regression
  - standardisierte Regressionskoeffizienten
  - Regression mit Dummyvariablen

Aufgabe 1: Fernsehkonsum und BMI (fiktiv)

Ein Gesundheitswissenschaftler interessiert sich dafür, ob die tägliche Fernsehdauer in Minuten einen Einfluss auf den Body-Mass-Index eines Befragten hat. Für fünf Befrage erhält er folgende Werte:

Person	1	2	3	4	5
Fernsehdauer in Minuten	0	60	120	150	240
Body-Mass-Index	22	18	25	30	25

a) Mit welchem Verfahren lässt sich dieser Zusammenhang untersuchen? Wie lautet das allgemeine Modell?

b) Wenden Sie das Verfahren an! Interpretieren Sie Ihr Ergebnis!

c) Wie gut ist die Fernsehdauer in Minuten zur Prognose des BMI geeignet? Interpretieren Sie Ihr Ergebnis.

Aufgabe 1a: Lösung

asymmetrische Fragestellung: Fernsehdauer (X)→ BMI (Y)

Ein Gesundheitswissenschaftler interessiert sich dafür, ob die tägliche Fernsehdauer in Minuten einen Einfluss auf den Body-Mass-Index eines Befragten hat.

zwei metrische Variablen

Person	1	2	3	4	5
Fernsehdauer in Minuten	0	60	120	150	240
Body-Mass-Index	22	18	25	30	25

bivariate lineare Regression

Modellgleichung: $y_i = b_0 + b_1 * x_i + e_i$

Aufgabe 1b: Lösung

ID	(x_i)	(y_i)	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x}) * (y_i - \bar{y})$
1	0	22					
2	60	18					
3	120	25					
4	150	30					
5	240	25					
	$\bar{x} =$	$\bar{y} =$		$SAQ_X =$		$SAQ_Y =$	$SP_{X,Y} =$

Aufgabe 1b: Lösung II

ID	(x_i)	(y_i)	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x}) * (y_i - \bar{y})$
1	0	22	-114	12996	-2	4	228
2	60	18	-54	2916	-6	36	324
3	120	25	6	36	1	1	6
4	150	30	36	1296	6	36	216
5	240	25	126	15876	1	1	126
	$\bar{x} = 114$	$\bar{y} = 24$		$SAQ_X = 33120$		$SAQ_Y = 78$	$SP_{X,Y} = 900$

Aufgabe 1b: Lösung III

- Berechnung Regressionsgewicht b_1:
  - $b_1 = \frac{SP_{X,Y}}{SAQ_X}$
  - $b_1 = \frac{900}{33120}$
  - $b_1 = 0{,}027 \ldots$
- Berechnung Regressionskonstante b_0:
  - $b_0 = \bar{y} - b_1 * \bar{x}$
  - $b_0 = 24 - 0{,}027 \ldots * 114$
  - $b_0 = 20{,}902$

- $\bar{x} = 114$
- $\bar{y} = 24$
- $SAQ_X = 33120$
- $SAQ_Y = 78$
- $SP_{X,Y} = 900$

Aufgabe 1b: Lösung IV

- Interpretation Regressionskonstante:
  - $b_0 = 20{,}902$
  - Die Regressionskonstante ist geometrisch gesehen der Schnittpunkt mit der y-Achse. Bei einem x-Wert von 0 würde hier folglich ein Wert von 20,902 auf der y-Achse vorhergesagt.
  - Inhaltlich bedeutet dies an dieser Stelle, dass für eine Person mit einer täglichen Fernsehdauer von 0 Minuten ein BMI von 20,902 vorhergesagt würde.

Aufgabe 1b: Lösung V

- Interpretation Regressionsgewicht:
  - $b_1 = 0{,}027$ ...
  - Dies ist geometrisch gesehen der Steigungskoeffizient. Wenn sich x um eine Einheit erhöht würde eine Erhöhung von y um 0,027 Einheiten vorhergesagt.
  - Dies bedeutet, wenn mit jeder zusätzlichen Minute Fernsehdauer eine Erhöhung des BMI-Wert um 0,027 Punkte vorhergesagt werden würde.

Aufgabe 1c: Lösung

- Berechnung der Güte des Modells:
  - bivariate Regression: $R^2 = (r_{X,Y})^2$
  - $r_{X,Y} = \frac{SP_{X,Y}}{\sqrt{SAQ_X * SAQ_Y}}$
  - $r_{X,Y} = \frac{900}{\sqrt{33120 * 78}}$
  - $r_{X,Y} \approx 0{,}560$
  - $R^2 = (r_{X,Y})^2 = 0{,}314$
- Interpretation:
  - Durch Kenntnis der Fernsehdauer der Befragten lässt sich die Prognose des BMI um 31,4% verbessern. (Hier fiktiv, in der Praxis natürlich deutlich geringer)

- $\bar{x} = 114$
- $\bar{y} = 24$
- $SAQ_X = 33120$
- $SAQ_Y = 78$
- $SP_{X,Y} = 900$

Aufgabe 1: SPSS-Output

Modellübersicht

Modell	R	R-Quadrat	Angepasstes R-Quadrat	Standardfehler der Schätzung
1	,560[a]	,314	,085	4,22467

a. Prädiktoren: (Konstante), minuten

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten	
		B	Standardfehler
1	(Konstante)	20,902	3,252
	minuten	,027	,023

a. Abhängige Variable: bmi

Aufgabe 2: Fernsehdauer und Gewicht II

Der Gesundheitswissenschaftler von gerade eben interessiert sich immer noch für den Einfluss der täglichen Fernsehdauer auf den BMI. Anstelle der täglichen Zeit Minuten verwendet er jedoch nun die tägliche Zeit in Stunden (Stunden=Minuten/60). Hierbei entsteht folgende Ausgangstabelle:

Person	1	2	3	4	5
Fernsehdauer in Stunden	0	1	2	2,5	4
Body-Mass-Index	22	18	25	30	25

a) Bestimmen Sie die entsprechende Regressionsgleichung. Wie verändern sich die Regressionskonstante und das Regressionsgewicht im Vergleich zum vorherigen Modell?

b) Überlegen Sie, wie sich Pearsons $r_{X,Y}$ und der Determinationskoeffizient R^2 verändern werden. Überprüfen Sie Ihre Vermutung!

Aufgabe 2a: Lösung I

ID	(x_i)	(y_i)	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x}) * (y_i - \bar{y})$
1	0	22			-2	4	
2	1	18			-6	36	
3	2	25			1	1	
4	2,5	30			6	36	
5	4	25			1	1	
	$\bar{x} =$	$\bar{y} = 24$				$SAQ_Y = 78$	$SP_{X,Y} =$

Aufgabe 2a: Lösung II

ID	(x_i)	(y_i)	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x}) * (y_i - \bar{y})$
1	0	22	-1,9	3,61	-2	4	3,8
2	1	18	-0,9	0,81	-6	36	5,4
3	2	25	0,1	0,01	1	1	0,1
4	2,5	30	0,6	0,36	6	36	3,6
5	4	25	2,1	4,41	1	1	2,1
	$\bar{x} = 1{,}9$	$\bar{y} = 24$		$SAQ_X = 9{,}2$		$SAQ_Y = 78$	$SP_{X,Y} = 15$

Aufgabe 2b: Lösung III

- Berechnung Regressionsgewicht b_1:
  - $b_1 = \frac{SP_{X,Y}}{SAQ_X}$
  - $b_1 = \frac{15}{9{,}2}$
  - $b_1 = 1{,}630$
- Berechnung Regressionskonstante b_0:
  - $b_0 = \bar{y} - b_1 * \bar{x}$
  - $b_0 = 24 - 1{,}630 \ldots * 1{,}9$
  - $b_0 = 20{,}902$
- Vergleich:
  - Regressionskonstante bleibt gleich, aber Regressionsgewicht ist 60x so groß.

- $\bar{x} = 1{,}9$
- $\bar{y} = 24$
- $SAQ_X = 9{,}2$
- $SAQ_Y = 78$
- $SP_{X,Y} = 15$

Aufgabe 2c: Lösung

- Berechnung der Güte des Modells:
  - bivariate Regression: $R^2 = (r_{X,Y})^2$
  - $r_{X,Y} = \frac{SP_{X,Y}}{\sqrt{SAQ_X * SAQ_Y}}$
  - $r_{X,Y} = \frac{15}{\sqrt{9{,}2 * 78}}$
  - $r_{X,Y} \approx 0{,}560$
  - $R^2 = (r_{X,Y})^2 = 0{,}314$
- Vergleich:
  - Die Erklärungskraft der beiden Modelle ist gleich.

- $\bar{x} = 1{,}9$
- $\bar{y} = 24$
- $SAQ_X = 9{,}2$
- $SAQ_Y = 78$
- $SP_{X,Y} = 15$

Aufgabe 2: Output SPSS

Modellübersicht

Modell	R	R-Quadrat	Angepasstes R-Quadrat	Standardfehler der Schätzung
1	,560[a]	,314	,085	4,22467

a. Prädiktoren: (Konstante), stunden

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten	
		B	Standardfehler
1	(Konstante)	20,902	3,252
	stunden	1,630	1,393

a. Abhängige Variable: bmi

Unstandardisiertes Regressionsgewicht b_1: Vorteile und Nachteile

- Vorteil:
  - lässt sich inhaltlich relativ intuitiv interpretieren: „Anstieg von X um eine Einheit bewirkt im Durchschnitt von Y um b_1 Einheiten“
  - eignet sich zum Aufstellen der Vorhersagegleichung
- Nachteil:
  - Höhe ist immer abhängig von den Maßeinheiten der beteiligten Variablen
  - kein „standardisiertes Zusammenhangsmaß“
  - besonders im multiplen Modell: misst nur die Stärke des absoluten Einflusses, aber nicht des relativen Einflusses zweier Variablen

Standardisiertes Regressionsgewicht b_1^*

- Ziel:
  - Einfluss der Maßeinheiten entfernen
- 1. Möglichkeit:
  - Durchführung einer z-Transformation (Standardisierung) der beteiligten X- und Y-Variablen
  - anschließend reguläre Bestimmung der Regressionsgleichung
- 2. Möglichkeit:
  - Standardisierung des Regressionsgewichts im Nachhinein

Z-Transformation (Standardisierung)

- Berechnung:
  - $z_i = \frac{x_i - \bar{x}}{s_X}$
- Inhaltliche Bedeutung:
  - 1. Schritt: Variable x wird zunächst an ihrem Mittelwert zentriert, dadurch erhält die neue Variable z einen Mittelwert von 0
  - 2. Schritt: Normierung anhand der Standardabweichung von x, dadurch hat die neue Variable z eine Standardabweichung bzw. Varianz von 1

Aufgabe 3: Fernsehdauer und BMI

Führen Sie eine z-Transformation der x-Variablen Fernsehdauer in Stunden durch.

Person	1	2	3	4	5
Fernsehdauer in Stunden	0	1	2	2,5	4
Body-Mass-Index	22	18	25	30	25

Hinweis:

- $z_i = \frac{x_i - \bar{x}}{s_X}$
- $s_X = \sqrt{\frac{SAQ_X}{n}}$

Aufgabe 3: Lösung

ID	(x_i)	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$z_{x_i} = \frac{x_i - \bar{x}}{s_X}$
1	0	-1,9	3,61	
2	1	-0,9	0,81	
3	2	0,1	0,01	
4	2,5	0,6	0,36	
5	4	2,1	4,41	
	$\bar{x} = 1{,}9$		$SAQ_X = 9{,}2$	

- $s_X = \sqrt{\frac{SAQ_X}{n}}$
- $s_X = \sqrt{\frac{9{,}2}{5}}$
- $s_X = 1{,}3564 \ldots$

Aufgabe 3: Lösung II

ID	(x_i)	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$z_{x_i} = \frac{x_i - \bar{x}}{s_X}$
1	0	-1,9	3,61	-1,4007
2	1	-0,9	0,81	-0,6635
3	2	0,1	0,01	0,0737
4	2,5	0,6	0,36	0,4423
5	4	2,1	4,41	1,5481
	$\bar{x} = 1{,}9$		$SAQ_X = 9{,}2$	
			$s_X = 1{,}3564$	

- $s_X = \sqrt{\frac{SAQ_X}{n}}$
- $s_X = \sqrt{\frac{9{,}2}{5}}$
- $s_X = 1{,}3564 \ldots$

Standardisiertes Regressionsgewicht b_1^* : über Standardisierung Ausgangsvariablen I

ID	(y_i)	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$z_{y_i} = \frac{y_i - \bar{y}}{s_Y}$
1	22	-2	4	-0,5064
2	18	-6	36	-1,5191
3	25	1	1	0,2532
4	30	6	36	1,5191
5	25	1	1	0,2532
	$\bar{y} = 24$		$SAQ_Y = 78$	
			$s_y = 3{,}9497$	

Standardisiertes Regressionsgewicht b_1^* : über Standardisierung Ausgangsvariablen II

Person	1	2	3	4	5
Fernsehdauer in Minuten	0	1	2	2,5	4
Body-Mass-Index	22	18	25	30	25

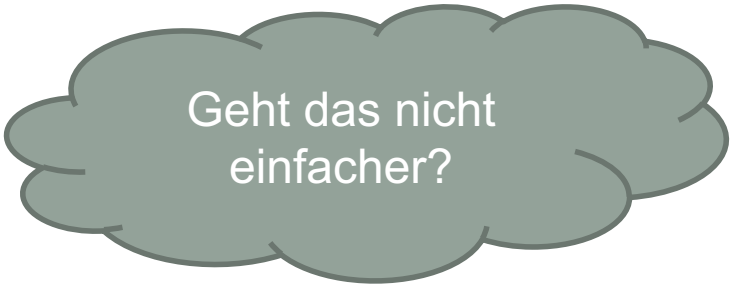
Z-Transformation

Person	1	2	3	4	5
standard. Fernsehdauer	-1,4007	-0,6635	0,0737	0,4423	1,5481
standard. Body-Mass-Index	-0,5064	-1,5191	0,2532	1,5191	0,2532

Standardisiertes Regressionsgewicht b_1^*: über Standardisierung Ausgangsvariablen IV

ID	(x_i)	(y_i)	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x}) * (y_i - \bar{y})$
1	-1,4007	-0,5064	-1,4	1,962	-0,5064	0,256	0,709
2	-0,6635	-1,5191	-0,66	0,44	-1,5191	2,308	1,008
3	0,0737	0,2532	0,074	0,005	0,2532	0,064	0,019
4	0,4423	1,5191	0,442	0,196	1,5191	2,308	0,672
5	1,5481	0,2532	1,548	2,397	0,2532	0,064	0,392
	$\bar{x} = 0$	$\bar{y} = 0$		$SAQ_X = 5$		$SAQ_Y = 5$	$SP_{X,Y} = 2{,}8$

- $b_1^* = \frac{SP_{X,Y}}{SAQ_X} = \frac{2{,}8}{5} = 0{,}56$
- $b_0^* = \bar{y} - b_1 * \bar{x} = 0 - 0{,}56 * 0 = 0$



Hinweis: Hier sehen wir auch, dass durch die Standardisierung beide Variablen einen Mittelwert von 0 und eine Varianz / Standardabweichung von 1 haben.

Standardisiertes Regressionsgewicht b_1^* : über nachträgliches Standardisieren

- Berechnung:
  - $b_1^* = b_1 * \frac{s_X}{s_Y}$
  - $b_1^* = r_{X,Y}$

s_X: Standardabweichung von X
s_Y: Standardabweichung von Y
b_1: unstandardisiertes Regressionsgewicht
$r_{X,Y}$: Pearsons Produktmomentkorrelation

- Interpretation:
  - Wenn sich x um eine Standardabweichung erhöht, so erhöht sich y im Durchschnitt um b_1^* Standardabweichungen
  - im bivariaten Fall zusätzlich wie $r_{X,Y}$ interpretierbar, da identisch
  - im multiplen Modell Maß der relativen Einflussstärke (welche Variable hat einen stärken relativen Einfluss?)

Aufgabe 4: Fernsehdauer und Gewicht II

Der Gesundheitswissenschaftler von gerade eben interessiert sich immer noch für den Einfluss der täglichen Fernsehdauer auf den BMI. Mithilfe von SPSS erhält er folgendes Regressionsmodell:

Modell		Nicht standardisierte Koeffizienten	
		B	Standardfehler
1	(Konstante)	20,902	3,252
	stunden	1,630	1,393

a. Abhängige Variable: bmi

Darüber hinaus ist bekannt, dass die Standardabweichung für die Fernsehdauer $s_X = 1{,}3565$ Stunden beträgt und die Standardabweichung für den BMI $s_Y = 3{,}9497$ ist.

a) Bestimmen Sie das standardisierte Regressionsgewicht für die Fernsehdauer.

b) Interpretieren Sie Ihr Ergebnis.

Aufgabe 4a: Lösung

- gegeben:
  - $b_0 = 20{,}902$
  - $b_1 = 1{,}630$
  - $s_X = 1{,}3565$
  - $s_y = 3{,}9497$
- gesucht:
  - b_1^*
- Berechnung:
  - $b_1^* = b_1 * \frac{s_X}{s_Y}$
  - $b_1^* = 1{,}630 * \frac{1{,}3565}{3{,}9497}$
  - $b_1^* = 0{,}560$

Modell		Nicht standardisierte Koeffizienten	
		B	Standardfehler
1	(Konstante)	20,902	3,252
	stunden	1,630	1,393

a. Abhängige Variable: bmi

Aufgabe 4b: Lösung

- Interpretation:
  - $b_1^* = 0{,}560$
  - Wenn x um eine Standardabweichung steigt so erhöht sich y im Durchschnitt um 0,560 Standardabweichungen.
  - Dies bedeutet, dass wenn der Fernsehkonsum um eine Standardabweichung ansteigt, sich der BMI um 0,56 Standardabweichungen erhöht.
  - Da im bivariaten Fall der Wert mit Pearsons $r_{X,Y}$ identisch ist, lässt sich auch sagen, dass ein hoher positiver Zusammenhang zwischen dem Fernsehkonsum und dem BMI-Wert besteht.

SPSS-Output

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	t	Sig.
		B	Standardfehler	Beta		
1	(Konstante)	20,902	3,252		6,428	,008
	stunden	1,630	1,393	,560	1,171	,326

a. Abhängige Variable: bmi

unstandardisiertes Regressionsgewicht b_1

standardisiertes Regressionsgewicht b_1^*

Vergleich beider Modelle

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	t	Sig.
		B	Standardfehler	Beta		
1	(Konstante)	20,902	3,252		6,428	,008
	stunden	1,630	1,393	,560	1,171	,326

a. Abhängige Variable: bmi

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	t	Sig.
		B	Standardfehler	Beta		
1	(Konstante)	20,902	3,252		6,428	,008
	minuten	,027	,023	,560	1,171	,326

a. Abhängige Variable: bmi

Vergleich beider Modelle II

ANOVA[a]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Regression	24,457	1	24,457	1,370	,326[b]
	Residuum	53,543	3	17,848		
	Gesamtsumme	78,000	4			

a. Abhängige Variable: bmi

b. Prädiktoren: (Konstante), stunden

ANOVA[a]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Regression	24,457	1	24,457	1,370	,326[b]
	Residuum	53,543	3	17,848		
	Gesamtsumme	78,000	4			

a. Abhängige Variable: bmi

b. Prädiktoren: (Konstante), minuten

Vergleich beider Modelle III

Modellübersicht

Modell	R	R-Quadrat	Angepasstes R-Quadrat	Standardfehler der Schätzung
1	,560[a]	,314	,085	4,22467

a. Prädiktoren: (Konstante), stunden

Modellübersicht

Modell	R	R-Quadrat	Angepasstes R-Quadrat	Standardfehler der Schätzung
1	,560[a]	,314	,085	4,22467

a. Prädiktoren: (Konstante), minuten

Bivariate Dummy-Regression

- Dummyvariablen:
  - Variablen mit ausschließlich zwei Ausprägungen (0/1)
  - Die Kategorie 0 wird als Referenzkategorie bezeichnet
  - i.d.R. steht 0 für das Nichtvorliegen eines Merkmals und 1 für das vorliegen eines Merkmals wenn möglich
- Dummyvariablen in der Regression
  - neben metrischen Variablen können auch Dummyvariablen als unabhängige Variablen (=X-Variablen!) verwendet werden
  - bei der Interpretation sind einige Besonderheiten zu beachten

Bivariate Dummy-Regression II

- Regressionsgleichung:
  - $y_i = b_0 + b_1 * x_i + e_i$ oder
  - $Y = b_0 + b_1 * D + e$
  - x_i bzw. D kann nur den Wert 0 oder 1 annehmen
- Prognosegleichung:
  - $\hat{y}_i = b_0 + b_1 * x_i$
  - $\hat{Y} = b_0 + b_1 * D$

Bivariate Dummy-Regression III

- Prognosegleichung für Referenzkategorie 0
  - $\hat{y}_i = b_0 + b_1 * 0$
  - $\hat{y}_i = b_0$
    → Regressionskonstante gibt Vorhersagewert von Y für Referenzkategorie an
- Prognosegleichung für Kategorie 1
  - $\hat{y}_i = b_0 + b_1 * 1$
  - $\hat{y}_i = b_0 + b_1$
    → Regressionskonstante b_0 plus Regressionsgewicht b_1 gibt Vorhersagewert von Y für die Kategorie mit der Ausprägung 1 an
    → Regressionsgewicht b_1 gibt Erhöhung im Vergleich zur Referenzkategorie 0 an, wenn die Ausprägung auf 1 steigt

Bivariate Dummy-Regression IV: Interpretation

- Regressionskonstante b_0:
  - gibt den vorhergesagten Wert von y für die Referenzkategorie x=0 an
- unstandardisiertes Regressionsgewicht b_1:
  - gibt die Erhöhung an, wenn sich die Ausprägung von x=0 auf x=1 ändert
- standardisiertes Regressionsgewicht b_1^*:
  - ACHTUNG: bei Dummyvariablen nicht interpretierbar!

Bivariate Dummy-Regression III: Beispiel

Eine Soziologin interessiert sich dafür, welchen Einfluss das Geschlecht (0=Frau, 1=Mann) auf das Nettoeinkommen eines Befragten hat. Mithilfe des ALLBUS 2014 erhält sie folgende Tabelle:

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	t	Sig.
		B	Standardfehler	Beta		
1	(Konstante)	1099,014	37,756		29,108	,000
	Geschlecht (rekodiert)	789,676	52,958	,253	14,911	,000

a. Abhängige Variable: BFR.:NETTOEINKOMMEN<OFFENE+LISTENANGABE>

Interpretieren Sie das Ergebnis inhaltlich!

Bivariate Dummy-Regression III: Beispiel

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	t	Sig.
		B	Standardfehler	Beta		
1	(Konstante)	1099,014	37,756		29,108	,000
	Geschlecht (rekodiert)	789,676	52,958	,253	14,911	,000

a. Abhängige Variable: BFR.:NETTOEINKOMMEN<OFFENE+LISTENANGABE>

- Im vorliegenden Fall ist die Referenzkategorie Frau. Da hier die Konstante b_0 bei 1099,014 liegt, bedeutet dies, dass für Frauen ein durchschnittliches Nettoeinkommen von 1099,01 Euro haben.
- Das Regressionsgewicht b_1 für das Geschlecht liegt bei 789,676. Wenn sich die Ausprägung der Variablen X von 0 auf 1 ändert, wird ein um etwa 789,676 höheres Einkommen vorhergesagt. Dies bedeutet hier, dass Männer mit 789,676 Euro mehr als Frauen rechnen können. Sie haben also im Durchschnitt ein Einkommen von 1888,69 Euro.

Aufgabe 5: Geschlecht und Nettoeinkommen

Für das soeben aufgestellte Modell Einfluss des Geschlechts soll die Güte des Modells bestimmt werden. Leider ist der SPSS-Output unvollständig und sie haben nur die ANOVA-Tabelle zur Verfügung:

ANOVA[a]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Regression	507149404,0	1	507149404,0	222,352	,000[b]
	Residuum	7417288502	3252	2280839,023		
	Gesamtsumme	7924437906	3253			

a. Abhängige Variable: BFR.:NETTOEINKOMMEN<OFFENE+LISTENANGABE>

b. Prädiktoren: (Konstante), Geschlecht (rekodiert)

a) Bestimmen Sie den Determinationskoeffizienten R^2.

b) Interpretieren Sie Ihr Ergebnis inhaltlich!

Aufgabe 5: Lösung

ANOVA[a]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Regression	507149404,0	1	507149404,0	222,352	,000[b]
	Residuum	7417288502	3252	2280839,023		
	Gesamtsumme	7924437906	3253			

a. Abhängige Variable: BFR.:NETTOEINKOMMEN<OFFENE+LISTENANGABE>

b. Prädiktoren: (Konstante), Geschlecht (rekodiert)

- Berechnung:
  - $R^2 = \frac{E_0 - E_1}{E_0} = \frac{SAQ_{Regresion}}{SAQ_{Gesamt}}$
  - $R^2 = \frac{507149404}{7924437906}$
  - $R^2 = 0{,}064$

Aufgabe 5: Lösung II

Modellübersicht

Modell	R	R-Quadrat	Angepasstes R-Quadrat	Standardfehler der Schätzung
1	,253[a]	,064	,064	1510,245

a. Prädiktoren: (Konstante), Geschlecht (rekodiert)

- Interpretation:
  - $R^2 = 0{,}064$
  - Durch Kenntnis von X lässt sich die Prognose von Y um 6,4% verbessern. Dies bedeutet hier, dass sich durch Kenntnis des Geschlechts die Prognose des Nettoeinkommens um 6,4 % verbessert.
  - Alternativ könnte man sagen, dass durch das Modell 6,4% der Streuung gebunden („erklärt“) werden. 6,4% der Streuung des Einkommens lassen sich also mit Kenntnis von X „erklären“.

Aufgabe 6.1: Öffentlicher Dienst und Nettoeinkommen

Eine Wirtschaftswissenschaftlerin interessiert sich dafür, ob die eine Tätigkeit im öffentlichen Dienst (X=1)ein höheres Nettoeinkommen produziert als die Tätigkeit außerhalb des öffentlichen Dienstes (X=0). Mithilfe des ALLBUS 2014 erhält sie folgende Tabelle:

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	t	Sig.
		B	Standardfehler	Beta		
1	(Konstante)	1774,603	32,987		53,798	,000
	im öffentlichen Dienst (rek)	286,682	65,322	,109	4,389	,000

a. Abhängige Variable: BFR.:NETTOEINKOMMEN<OFFENE+LISTENANGABE>

Interpretieren Sie das Ergebnis inhaltlich!

Aufgabe 6.1: Lösung

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	t	Sig.
		B	Standardfehler	Beta		
1	(Konstante)	1774,603	32,987		53,798	,000
	im öffentlichen Dienst (rek)	286,682	65,322	,109	4,389	,000

a. Abhängige Variable: BFR.:NETTOEINKOMMEN<OFFENE+LISTENANGABE>

- Regressionskonstante b_0:
  - Die Referenzkategorie, also die Befragten außerhalb des öffentlichen Dienstes, erhalten etwa ein durchschnittliches Nettoeinkommen von 1774,60 Euro.
- Regressionsgewicht b_1:
  - Demgegenüber erhalten beschäftigte im öffentlichen Dienst etwa 286,68 Euro mehr, also im Durchschnitt etwa 2061,29 Euro.

Aufgabe 6.2: Öffentlicher Dienst und Nettoeinkommen II

Eine Wirtschaftswissenschaftlerin interessiert sich dafür, ob die eine Tätigkeit im öffentlichen Dienst (X=1)ein höheres Nettoeinkommen produziert als die Tätigkeit als die Tätigkeit außerhalb des öffentlichen Dienstes (X=0). Mithilfe des ALLBUS 2014 erhält sie folgende Tabelle:

ANOVA[a]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Regression	24919778,70	1	24919778,70	19,261	,000[b]
	Residuum	2062283446	1594	1293778,824		
	Gesamtsumme	2087203224	1595			

a. Abhängige Variable: BFR.:NETTOEINKOMMEN<OFFENE+LISTENANGABE>

b. Prädiktoren: (Konstante), im öffentlichen Dienst (rek)

Wie gut eignet sich die Tätigkeit im öffentlichen Dienst zur Prognose des Einkommens?

Aufgabe 6.2: Lösung

ANOVA[a]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Regression	24919778,70	1	24919778,70	19,261	,000[b]
	Residuum	2062283446	1594	1293778,824		
	Gesamtsumme	2087203224	1595			

a. Abhängige Variable: BFR.:NETTOEINKOMMEN<OFFENE+LISTENANGABE>

b. Prädiktoren: (Konstante), im öffentlichen Dienst (rek)

- Berechnung:
  - $R^2 = \frac{E_0 - E_1}{E_0} = \frac{SAQ_{Regresion}}{SAQ_{Gesamt}}$
  - $R^2 = \frac{24919778{,}7}{2062283446}$
  - $R^2 = 0{,}012$
- Interpretation:
  - Durch Kenntnis, ob ein Befragten im öffentlichen Dienst tätig ist, verbessert sich die Vorhersage lediglich um 1,2%. Das ist nicht viel.

Literaturhinweise

- Wie letzte Woche
- Zusätzlich:
  Kerstin Völkl / Christoph Korb (2018): Deskriptive Statistik. Eine Einführung für Politikwissenschaftlerinnen und Politikwissenschaftler. S. 230-234.

Übungsaufgabe 1: Geschlecht und Gewicht

Eine Gesundheitsforscherin interessiert sich dafür, ob sich Männer und Frauen hinsichtlich ihres Gewichtes unterscheiden. Das Geschlecht ist dabei dummycodiert, wobei der Wert 0 für Frauen und der Wert 1 für Männer steht. Mithilfe des ALLBUS 2014 ergibt sich folgende Prognosegleichung: $\hat{y} = 70{,}166 + 15{,}135 * D$.

a) Interpretieren Sie die Regressionskonstante b_0 und das Regressionsgewicht b_1! (2)
b) Welches Gewicht würde mithilfe der Gleichung für einen Mann vorhergesagt? (1)
c) Wie würde die Prognosegleichung aussehen, wenn bei der Codierung der Dummyvariablen Männer und Frauen getauscht werden, also Männer den Wert 0 hätten und Frauen den Wert 1? (3)

Übungsaufgabe 1a: Lösung

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
		B	Standardfehler	Beta	t	Sig.
1	(Konstante)	70,166	,362		193,983	,000
	Geschlecht (rek)	15,135	,505	,456	29,978	,000

a. Abhängige Variable: GEWICHT IN KG, BEFRAGTE<R>

- Die Regressionskonstante b_0 liegt bei 70,166. Dies bedeutet, dass die angehörigen der Referenzkategorie Frau (D=0) ein durchschnittliches Gewicht von 70,166 kg haben. ($\hat{y}_{Frau} = 70{,}166 + 15{,}135 * 0$).
- Das Regressionsgewicht b_1 liegt bei 15,135. Die bedeutet, dass die Männer (D=1) im Durchschnitt 15,135 kg mehr wiegen als die Frauen.

Übungsaufgabe 1b: Lösung

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	t	Sig.
		B	Standardfehler	Beta		
1	(Konstante)	70,166	,362		193,983	,000
	Geschlecht (rek)	15,135	,505	,456	29,978	,000

a. Abhängige Variable: GEWICHT IN KG, BEFRAGTE<R>

- Die Prognosegleichung für die Männer würde die folgt aussehen: $\hat{y}_{Mann} = 70{,}166 + 15{,}135 * 1 = 85{,}301$.
- Das bedeutet das Männer im Durchschnitt 85,301 kg wiegen.

Übungsaufgabe 1c: Lösung

- Wenn wir die Codierung tauschen, sind die Männer nun die Referenzkategorie (D=0).
- Da wir wissen aus der vorherigen Aufgabe wissen, dass Männer im Durchschnitt 85,301 kg wiegen, ist unsere neue Regressionskonstante $b_0 = 85{,}301$.
- Darüber hinaus ist bekannt, das Frauen im Durchschnitt 15,135 kg weniger wiegen als Männer. Folglich ist unser Regressionsgewicht $b_1 = -15{,}135$.
- Das Modell lautet dementsprechend wie folgt: $\hat{y} = 85{,}301 - 15{,}135 * D$.

Übungsaufgabe 1c: Lösung II

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	t	Sig.
		B	Standardfehler	Beta		
1	(Konstante)	85,301	,352		242,183	,000
	Geschlecht (rek2)	-15,135	,505	-,456	-29,978	,000

a. Abhängige Variable: GEWICHT IN KG, BEFRAGTE<R>

Übungsaufgabe 2: Geschlecht und Gewicht

Mithilfe von SPSS erhält der Gesundheitswissenschaftler von gerade eben folgenden ANOVA-Output:

ANOVA[a]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Regression	195769,040	1	195769,040	898,664	,000[b]
	Residuum	744810,554	3419	217,845		
	Gesamtsumme	940579,594	3420			

a. Abhängige Variable: GEWICHT IN KG, BEFRAGTE<R>

b. Prädiktoren: (Konstante), Geschlecht (rek2)

Wie gut eignet sich das Geschlecht zur Prognose des Gewichts eines Befragten? Interpretieren Sie Ihr Ergebnis inhaltlich und statisch (4)

Übungsaufgabe 2: Lösung

ANOVA[a]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Regression	195769,040	1	195769,040	898,664	,000[b]
	Residuum	744810,554	3419	217,845		
	Gesamtsumme	940579,594	3420			

a. Abhängige Variable: GEWICHT IN KG, BEFRAGTE<R>

b. Prädiktoren: (Konstante), Geschlecht (rek2)

- Berechnung:
  - $R^2 = \frac{E_0 - E_1}{E_0}$ (1)
  - $R^2 = \frac{195769{,}040}{940579{,}594}$ (1)
  - $R^2 = 0{,}208$ (1)
- Interpretation:
  - Durch Kenntnis des Geschlechts lässt sich die Prognose des Gewichts um 20,8 % verbessern. (1)

Übungsaufgabe 3: US-Internetnutzung

Eine US-Mediensoziologin interessiert sich dafür, ob es einen Einfluss des Alters (in Jahren) eines Befragten auf seine Internetnutzung (in Stunden) gibt. Für die 1018 Befragten mit gültigen Werten bei beiden Variablen ergab sich mithilfe des General Social Surveys 2012 folgender SPSS-Output:

Modell		Nicht standardisierte Koeffizienten	
		B	Standardfehler
1	(Konstante)	18,470	1,292
	AGE OF RESPONDENT	-,182	,028

a. Abhängige Variable: WWW HOURS PER WEEK

Darüber hinaus ist bekannt, dass $\sum(x_i - \bar{x})^2 = 265344$ und $\sum(y_i - \bar{y})^2 = 214545$ liegt.

a) Stellen Sie die stochastische Regressionsgleichung auf!

b) Interpretieren Sie die Regressionskoeffizienten statistisch und inhaltlich.

c) Berechnen Sie das standardisierte Regressionsgewicht für das Alter und interpretieren Sie ihr Ergebnis!

d) Was würde mit dem standardisierten Regressionsgewicht passiert, wenn das Alter in Tagen statt in Jahren gemessen würde?

e) Berechnen Sie R^2 und interpretieren Sie dieses Maß!

Übungsaufgabe 3a und b: Lösung

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten	
		B	Standardfehler
1	(Konstante)	18,470	1,292
	AGE OF RESPONDENT	-,182	,028

a. Abhängige Variable: WWW HOURS PER WEEK

- Regressionsgleichung:
  - $y_i = 18{,}470 - 0{,}182 * x_i + e_i$
- Interpretation:
  - Für eine Person mit einem Alter von 0 Jahren wird im Durchschnitt ein wöchentlicher Internetkonsum von 18,470 Stunden pro Woche vorhergesagt. (nicht sinnvoll)
  - Mit jedem Jahr, die eine Person älter wird, fällt die durchschnittliche Internetkonsum um 0,182 Stunden.

Übungsaufgabe 3c: Lösung

- gegeben:
  - $b_1 = -0{,}182$
  - $n = 1082$
  - $SAQ_X = \sum(x_i - \bar{x})^2 = 265344 \rightarrow s_X^2 = \frac{265344}{1082}$
  - $SAQ_Y = \sum(y_i - \bar{y})^2 = 214545 \rightarrow s_Y^2 = \frac{214545}{1082}$
- gesucht:
  - $b_1^* = b_0 * \frac{s_X}{s_Y} = -0{,}182 * \frac{\sqrt{\frac{265344}{1082}}}{\sqrt{\frac{214545}{1082}}} = -0{,}2024$

Übungsaufgabe 3c: Lösung II

- Interpretation standardisiertes Regressionsgewicht b_1^*:
  - $b_1^* = -0{,}2024$
  - Wenn x um eine Standardabweichung steigt, sinkt y nach diesem Modell um durchschnittlich 0,2024 Standardabweichungen. Steigt also das Alter um eine Standardabweichung, so sinkt der Internetkonsum im Durchschnitt 0,2024 Standardabweichungen.
  - Im bivariaten Fall ist das standardisierte Regressionsgewicht identisch mit Pearsons Produktmomentkorrelationskoeffizient. Deswegen kann man sagen, dass ein schwacher negativer Zusammenhang zwischen dem Alter und dem Internetkonsum besteht.

Übungsaufgabe 3d: Lösung

- Interpretation:
  - Das standardisierte Regressionsgewicht ist unabhängig von den Maßeinheiten der Ausgangseinheiten. Aus diesem Grund würde sich das standardisierte Regressionsgewicht nicht verändern.
  - Lediglich das unstandardisierte Regressionsgewicht verändert sich.

Übungsaufgabe 3e: Lösung

- Berechnung:
  - $r_{X,Y} = b_1^*$
  - $r_{X,Y} = -0{,}2024$
  - $R^2 = (-0{,}2024)^2 = 0{,}0497$
- Interpretation:
  - Durch Kenntnis des Alters des Befragten lässt sich die Prognose des Internetkonsums um lediglich 4,97% verbessern.

Übungsaufgabe 3: SPSS

Modellübersicht

Modell	R	R-Quadrat	Angepasstes R-Quadrat	Standardfehler der Schätzung
1	,203[a]	,041	,040	14,230

a. Prädiktoren: (Konstante), AGE OF RESPONDENT

ANOVA[a]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Regression	8810,660	1	8810,660	43,511	,000[b]
	Residuum	205734,834	1016	202,495		
	Gesamtsumme	214545,494	1017			

a. Abhängige Variable: WWW HOURS PER WEEK

b. Prädiktoren: (Konstante), AGE OF RESPONDENT

Übungsaufgabe 3: SPSS II

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	t	Sig.
		B	Standardfehler	Beta		
1	(Konstante)	18,470	1,292		14,291	,000
	AGE OF RESPONDENT	-,182	,028	-,203	-6,596	,000

a. Abhängige Variable: WWW HOURS PER WEEK